A2. Project description

The Exploring Collaborations to Harness Objects in a Digital Environment for Preservation (ECHO DEPository) project aims to address the issues of how we collect, manage, preserve, and make useful the enormous amount of digital information our culture is now producing. Collecting, selecting and preserving digital information requires approaches and resources that are substantively different from those we have used traditionally.

The project is a partnership among the University of Illinois at Urbana-Champaign (the Library, the Graduate School of Library and Information Science, the National Center for Supercomputing Applications); the Online Computer Library Center (OCLC); the Perseus Project at Tuft's University; the Michigan State University Library; and an alliance of state libraries from Arizona, Connecticut, Illinois, North Carolina and Wisconsin. Key project activities will include articulating a methodology based on archival principles for selecting digital material for capture and preservation; building software tools based on this model to help automate the selection process; installing, configuring and testing existing open-source digital repository systems; and conducting research into the problems of long-term semantic preservation of digital resources. The tools, practices, evaluations and research findings that result from this partnership will aid institutions in selecting and preserving electronic resources in a variety of digital repositories, and provide insight into the continuing challenges of long-term digital preservation. Further details on each of the four key project activities are provided below.

1. Selection rationale

Articulating a rationale and methodology for selecting digital materials, whether web-accessible or not, as aggregates, rather than at the item level, based on archival principles, and using provenance, functional analysis, and context analysis to facilitate meta-tagging for retrieval.

This methodology differs from approaches utilized to date. Manual, item-level selection fails because information professionals cannot keep up with the enormous number of resources on the Web. A fully automated approach to capture all the Web results in substantive materials being buried under a mountain of ephemeral, redundant, or irrelevant information.

2. Tools development

Building software to facilitate selection and preservation of digital materials that is based on that methodology and that can be scaled for varying degrees of human intervention to complement the automated rules.

We will develop a harvesting tool to bridge the gap between manual selection and automated capture by transforming collecting policies into software-based rules and configurations. The tool will help

information professionals sort the wheat from the chaff, and add metadata to complement keyword searching. In the case of information that is not web-accessible or for which only limited derivative forms can be delivered to the web from a richer underlying source, batch-loading may be the most practical method.

3. Repository evaluation

Installing, configuring, and testing open-source and commercial digital repositories to evaluate the strengths and weaknesses of each with regard to types of content, user and uses, interoperability, implementation of standards, and technical requirements.

ii

The creation of digital repositories is in its infancy. The partners will implement at least one instance of each of the major digital repository architectures (DSpace, FEDORA, Greenstone, Eprints, the OCLC Digital Archive) and GSLIS faculty and other independent experts will evaluate their suitability for different types and amounts of content, different users and uses, and different administrative and technical environments. In particular, the partners will test the compatibility of ingest and dissemination across repositories to determine if any incompatibilities stem from differences in implementation of standards such as OAIS and METS or from immaturity of standards development.

4. Long-term semantic preservation research

Researching techniques to migrate the semantic content of documents (and document structures) across generations of encoding schemes.

Using sample content collected in the project, GSLIS researchers will carry out long-range research on techniques for migrating the semantic content of documents (and document structures) across generations of encoding schemes. Advances in automated markup interpretation and inference will be applied to the problems of long-term digital preservation.